Digitizing Premodern Text with the Chinese Text Project

Introduction

Since computers first became available to scholars working with premodern Chinese written materials, their potential utility has been widely explored. For some research purposes, simply having access to a digital surrogate of a research object in itself provides considerable practical advantages similar to those of possessing a physical copy – not needing to visit a physical library to examine a rare edition, for example. Transcribed and searchable digital texts additionally offer the transformative ability to locate desired passages containing particular words or phrases far more efficiently than would be possible with any printed copy, no matter how well indexed. Some benefits of the digital medium are attained as a direct consequence of digitization, because existing off-the-shelf software provides appropriate functionality; far greater benefits can be realized through the creation of digital systems designed specifically to work with these types of text. The Chinese Text Project (https://ctext.org/), first publicly released online in 2005, attempts to leverage opportunities offered by the digital medium to create a platform for working with premodern Chinese primary source materials conducive to their use in research and teaching. At present, this system includes over 25 million pages of primary source material contributed by university libraries and scanning centers, alongside digital transcriptions of these works, integrated and extensible tools for reading, searching, and navigating these materials, as well as tools and workflows for creating and editing digital transcriptions and performing a variety of computer-assisted text analysis and text mining tasks.

Digital representations of text

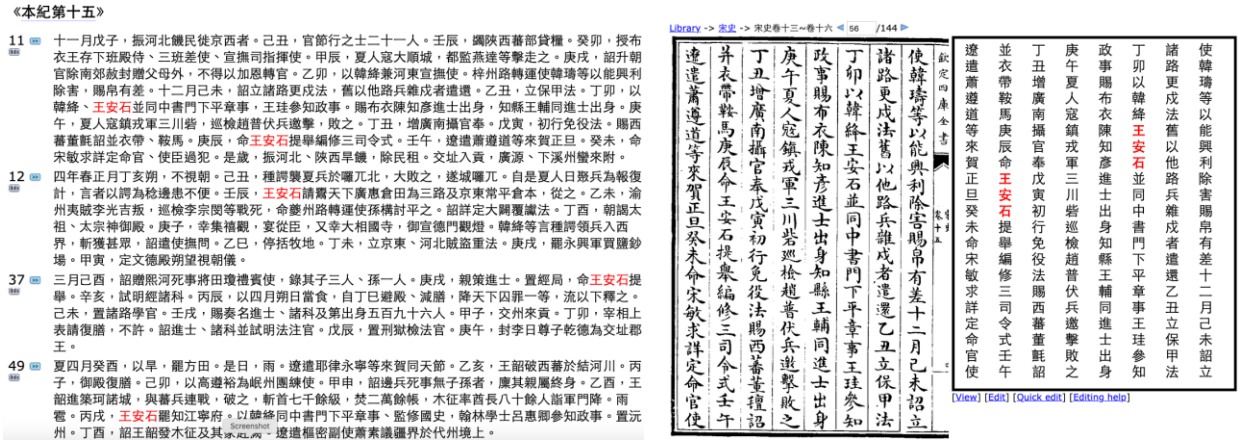
A fundamental challenge of representing textual material digitally is that different types of representation are often desirable in order to model different aspects of the content. The representation determines what can feasibly be done with the material computationally, and also directly affects the economic cost of producing a digitized version. Most fundamentally, there is the distinction between textual content modeled as images, describing precisely how each page of a particular text *appears,* versus the same pages of content modeled as digital text, containing a sequence of characters expressed abstractly using codes describing what the text *says* in a literal sense. The first of these is valuable in that it provides a precise record of the visually perceptible content of an historical object, and has the practical advantage of being created through a largely mechanical process.[1] This type of modeling creates a visual surrogate of the object, such that in principle almost any question that could be answered by visually inspecting the object itself can be answered instead by inspecting the surrogate. This offers huge benefits in terms of preservation of rare and unique objects, as well as greatly improved accessibility: seconds to call up and inspect an image, rather than minutes or hours to locate and consult the physical object directly. At the same time, a digitized image of a page of text contains no directly accessible representation of any of the letters appearing in the image; without access to this information, computer software can do relatively little with these page images beyond displaying them in sequence, zooming in and out on parts of the image, or offering to jump to a particular page by number. A textual representation of the same content makes possible processing using the letters of the language – most obviously full-text search.

---

[1] I do not mean here to trivialize the considerable effort and expertise which goes into professional digitization.

While it is possible for a computer to produce a textual representation from an image – this process being termed Optical Character Recognition (OCR) – unlike the initial digitization step, this is not a mechanical task, but a process performed by software attempting to mimic complex aspects of human cognition and understanding. Though much progress has been made, particularly when faced with additional challenges of historical documents errors are always to be expected as part of this process. An alternative is to have human beings type in the text; however, this is considerably more costly, and thus limits the scope of materials to which it can feasibly be applied.

Concerns about accuracy resulting from OCR and even manual input, combined with many other pitfalls in both the creation of digital transcriptions of text and their use for full-text search have made many researchers wary of reliance on full-text digital libraries. Some concerns can be mitigated by making use of both representations at the same time: using a textual representation to enable full-text search, while at the same time linking it to a visual surrogate of the original text. In the Chinese Text Project, this is realized by textual objects being presented through two separate "views" or interfaces (Figure 1). The first of these is a textual view, allowing navigation of the text according to its hierarchical structure, separated into fascicles, paragraphs, and other relevant subsections. The second view contains the same transcription rearranged so that it is matched page by page and column by column with images of the source document, navigable by the pages of the original source. Links allow switching between the two views, and full-text search results can be highlighted in both. These paired views form the basis of scholarly use of the system: while it is typically more efficient to work with the transcribed textual content, concerns about its accuracy and precision can be addressed by visual comparison of the relevant part of a transcription with the source edition.


《本紀第十五》

11 十一月戊子，振河北饑民徙京西者。己丑，官節行之士二十一人。壬辰，蠲陝西蕃部貸糧。癸卯，授布衣王存下班殿侍、三班差使、宣撫司指揮使。甲辰，夏人寇大順城，都監燕達等擊走之。庚戌，詔升朝官除南郊赦封贈父母外，不得以加恩轉官。乙卯，以韓絳兼河東宣撫使。梓州路轉運使韓璹等以能興利除害，賜帛有差。十二月己未，詔立諸路更戍法，舊以他路兵雜戍者遣還。乙丑，立保甲法。丁卯，以韓絳、王安石並同中書門下平章事，王珪參知政事。賜布衣陳知彥進士出身，知縣王輔同進士出身。庚午，夏人寇鎮戎軍三川砦，巡檢趙普伏兵邀擊，敗之。丁丑，增廣南攝官奉戊寅初行免役法。賜西蕃董氈詔並衣帶、鞍馬。庚辰，命王安石提舉編修三司令式。壬午，遼遣蕭遵道等來賀正旦。癸未，命宋敏求詳定命官、使臣過犯。是歲，振河北、陝西旱饑，除民租。交址入貢，廣源、下溪州蠻來附。

12 四年春正月丁亥朔，不視朝。己丑，種諤襲夏兵於囉兀北，大敗之，遂城囉兀。自是夏人日慕兵為報復計，言者以諤為稔邊患不便。壬辰，王安石請罷天下廣惠倉田為三路及京東常平倉本，從之。乙未，渝州夷賊李光吉叛，巡檢李宗閔等戰死，命夔州路轉運使孫構討平之。詔詳定大閱覆讞法。丁酉，朝謁太祖、太宗神御殿。庚子，幸集禧觀，宴從臣，又幸大相國寺，御宣德門觀燈。韓絳等言種諤領兵入西界，斬獲甚眾，詔遣使撫問。乙巳，停括牧地。丁未，立京東、河北賦盜重法。庚戌，罷永興軍買鹽鈔場。甲寅，定文德殿朔望視朝儀。

37 三月己酉，詔贈熙河死事將田瓊禮賓使，錄其子三人、孫一人。庚戌，親策進士。置經局，命王安石提舉。辛亥，試明經諸科。丙辰，以四月朔日當食，自丁巳避殿、減膳，降天下囚罪一等，流以下釋之。己未，置諸路學官。壬戌，賜秦名進士、諸科及第出身五百九十六人。甲子，交州來貢。丁卯，宰相上表請復膳，不許。詔進士、諸科並試明法注官。戊辰，置刑獄檢法官。庚午，封李日尊子乾德為交址郡王。

49 夏四月癸酉，以旱，罷方田。是日，雨。遼遣耶律永寧等來賀同天節。乙亥，王韶破西蕃於結河川。丙子，御殿復膳。己卯，以高遵裕為岷州團練使。甲申，詔邊兵死事無子孫者，廩其親屬終身。乙酉，王韶進築珂諸城，與蕃兵連戰，破之，斬首七千餘級，焚二萬餘帳，木征率首長八十餘人詣軍門降。雨雹。丙戌，王安石罷知江寧府。以韓絳同中書門下平章事、監修國史，翰林學士呂惠卿參知政事。置沅州。丁酉，詔王韶發木征及其家屬～遼遣樞密副使蕭素議疆界於代州境上。

Library -> 宋史 -> 宋史卷十三~卷十六 56 /144

欽定四庫全書 卷十五

使韓璹等以能興利除害賜帛有差十二月己未詔立諸路更戍法舊以他路兵雜戍者遣還乙丑立保甲法丁卯以韓絳王安石並同中書門下平章事王珪參知政事賜布衣陳知彥進士出身知縣王輔同進士出身庚午夏人寇鎮戎軍三川砦巡檢趙普伏兵邀擊敗之丁丑增廣南攝官奉戊寅初行免役法賜西蕃董氈詔並衣帶鞍馬庚辰命王安石提舉編修三司令式壬午遼遣蕭遵道等來賀正旦癸未命宋敏求詳定命官使

[View] [Edit] [Quick edit] [Editing help]

Figure 1: Full-text view (left) and image-and-transcription view (right) of the same search for "王安石" in part of a copy of the Songshi Songshi (宋史) in the Chinese Text Project.

Transcription, Crowdsourcing, and the "Long Tail"

Due to the difficulty of automatically turning an image of text into accurate computer-readable text, digital transcription of historical documents is a costly process. At the same time, the economic costs of image digitization have decreased significantly, leading to the production of large volumes of digitized images of historical primary sources – for premodern Chinese works, amounting to hundreds of millions of pages in total. Though the Chinese Text Project contains transcriptions of tens of thousands of texts which have been hand-typed and matched with their corresponding editions using

specially developed techniques,[2] there remains a substantial gap between the volume of historical material digitized photographically and the amount transcribed and publicly available.

In an attempt to deal with this challenge, a crowdsourcing approach is used to allow many individual users to gradually and collaboratively improve the quality of transcriptions created through OCR, with the ultimate goal of creating accurate digital editions. The open, public web interface – currently accessed by over 30,000 individual users every day – encourages users to correct errors in transcriptions directly from the image and transcription view shown in Figure 1. Users are able to directly edit the transcription of the page to correct errors, with the primary standard against which it should be corrected being the scanned page image of the same edition shown directly to beside the transcription.[3] Because transcriptions are aligned closely to the scanned images, identifying and correcting errors of transcription is a relatively straightforward task.

Much like in Wikipedia, version control is used to ensure that a precise log is maintained of all changes to a text, and to guarantee that in the event of a mistaken edit any text can be reverted to its state prior to that edit occurring. This log is available to all other users, and links to visual summaries of the precise changes made at any given time, as well as to the relevant page of text in the image and transcription view. This allows any other user to easily verify the accuracy of a submitted edit, resulting in a scalable, self-regulating system that does not rely on a traditional process of review. This approach has the benefit of making all texts available for use in the best possible state in which they exist at any given point in time. Where no manually input or corrected transcription is available, OCR-derived transcriptions are displayed and used to enable full-text search, and the user alerted to this fact; as soon as corrections are made, these improve the accuracy of the transcription and search function. This enables – in many cases for the first time – access to transcriptions of a "long tail" of less mainstream material, which has not previously attracted the attention of more traditional transcription projects.

Analysis and visualization of textual features

In addition to more efficient ways of accessing primary sources, digitization offers many new possibilities for working with these sources in ways that would simply not be practical in any other medium. Full-text search, on the scale of an individual work, might be thought of as nothing more than a faster means of locating information that might previously have been accessible through an index or concordance. However, as ever-growing amounts of material are digitized, even something as straightforward as keyword search can lead directly to new and transformative ways of working with text.

A natural extension to full-text search is its application at ever larger scales, facilitating the discovery of material in unknown or unexpected locations, as well as the confirmation of absence of appearance. In the Chinese Text Project, a "Global search" function facilitates this, simultaneously searching across all texts in the system, or across a subset of texts selected according to various metadata properties, and summarizing the results. While at first glance a straightforward-seeming task requiring nothing more than sufficient computational resources, useful summarization of the results generally relies upon additional information about the texts themselves. Without additional organizing principles, a search for a term occurring in hundreds of texts would simply produce a long, unordered

---

[2] Donald Sturgeon, "Large-scale Optical Character Recognition of Pre-modern Chinese Texts," *International Journal of Buddhist Thought and Culture* 28.2 (2018), 11–44.

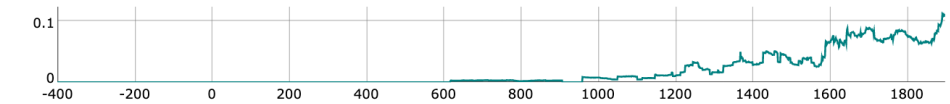
[3] Additional rules are used to cover more complex cases, such as instances of textual corruption in the edition being transcribed: https://ctext.org/instructions/wiki-formatting

list of occurrences, that might have to be examined one by one to yield any useful observations. At the scale of the Chinese Text Project – currently including over 30,000 texts and in excess of 5 billion characters of transcription, covering Warring States through Republican era texts – one of the most obvious variables to use in organizing this material is its approximate date of composition. This is particularly intuitive given the high rates of text reuse, quotation, etc. across the classical and premodern corpus: a chronologically ordered set of search results will naturally highlight the earliest attested occurrences of a term or phrase, which will in many (though by no means all) cases be the source of a particular saying, distinctive phrase, or even in some cases, word. Time also provides a meaningful principle for summarization and navigation when large numbers of results are generated, spanning hundreds or thousands of individual texts.

In order to facilitate this functionality, precise information about dates of composition for all texts is needed. While there are some straightforward cases, for many historical texts this is far from trivial to determine, and particularly for early texts also includes a high degree of uncertainty. Furthermore, many "texts" as modeled in full-text databases and digital libraries are in fact composite works of multiple authors – including a variety of common types of case such as commentaries, works with prefaces and postfaces, as well as explicitly multi-authored works – and thus a degree of imprecision is inherent in the task if the units to be assigned dates of authorship are entire *texts*.[4] While acknowledging the imprecision involved in this task, in the Chinese Text Project texts are assigned dates of authorship on the basis of the estimated dates of first being written down in substantially their present form, treating commentary (though not prefaces and postfaces) as part of the text. Dates are recorded as a range of years, representing uncertainty about year of composition. Additionally, texts that represent distinct editions of the same work, and thus have substantially identical contents, are recorded explicitly. This data is then used for two organizational purposes: to provide an ordering of all texts, such that (imprecision and errors notwithstanding) results in earlier texts appear before results in later texts without repetition of the same work, and to visualize the frequency with which terms or phrases occur in works over time. For searches with small numbers of matches in the corpus, this yields a short list of results, ordered by date of authorship; for more common terms, it provides an intuitive visual summary of how results occur in a sample of the written record across a timespan of more than 2000 years (Figure 2). In the current implementation, a chart offers a visual summary of the search results by plotting the proportion of all texts composed in any given year which contain the search term.[5] The chart both summarizes the information and provides an additional way of navigating it – selecting a span of years in the horizontal axis of the chart causes the specific instances from texts corresponding to that region of the chart to be displayed below.

[4] A logical enhancement of the approach described here is to further subdivide texts, using markup to explicitly record which parts are authored by which persons (if known) and during which time periods – recording the information that a preface is of different authorship to the main text, say, or that the text being commented upon predates the commentary.

[5] The contribution to this proportion of texts whose dates of authorship are imprecise is distributed equally across all years within the recorded range for that text.



Found in 574 texts.

唐	《白蓮集》：	諸徒弟時來弔石頭送陳霸歸閩凉風動行興含笑話臨途已得身名了全忘客道孤鄉程過百越帆影遶
北宋	《文苑英華》：	盃中過水鷗送日暮鍾交戌嶺叫雲寒角動城樓酒酣笑話秋風裏誰道槐花更起愁偶題林亭集作題友
	《雜纂二續》：	謬漢作文章。村裏人體段。好說脫空。好笑話人。還俗僧道舉止。愛說是非。貪財
	《龍川集》：	自致而聲音姓字之與通晚將歸休始獲見公握手一笑話言從容心滿意愜俯首來東三年之間竟安此窮人
	《西河集》：	必請曾晢仲弓聞不欲勿施必請之辟剛之所生此是笑話乃講師不通謂父兄長老之稱則諺云要好問三老
	《貴耳集》：	，無事早下山去。」託此為談謔之助，世人相傳笑話，餘因錄一二事以資好事者一笑。有知州未滿
	《四如集》：	已授涂君復之俾留東野壁間以為吾兒異日邂逅一笑話柄　　魚我亭記孟子曰萬物皆備於我則物物皆
	《五百家播芳大全文粹》：	不能自致而聲音之與通晚得歸休始獲見公握手一笑話言從容心痛意愜俯首來東三年之間竟安此窮人
南宋	《演山集》：	開道情有箇鶯鶯問歸舍楚山青臥影水天明松庵誰笑話見還嬰鶴歸日落聚兩忘情好笑人癡處白頭青塚
	《香溪集》：	留懷歸錫難駐長途雪没脛栗烈正寒沍披披拂風袖笑話巖中趣巾瓶到銅山鐘梵歲云暮浮雲亦何意南北
	《濟公活佛傳奇錄》：	請。」濟顛道：「倘若公書請不來時，卻要被我笑話，寺裏既無酒吃，我只得別尋主顧。」遂一逕
	《莊靖集》：	著回天力有客空揮卻日戈青眼乍驚行處少赤心常笑話中多愁人近亦無腸斷又聽離筵送別歌九日
	《江湖長翁集》：	為防市弩今朝河面吹細痕竹樹不聲人駿奔掣鈴一笑話疇昔便有樂事酬佳辰鎮淮主人開滌樽蕙蘭香前
	《莊靖先生遺集》：	著回天力有客空回卻日戈青眼乍驚行處少赤心常笑話中多愁人近亦無腸斷又聽離筵送別歌 ○九

Figure 2: Part of the first page of search results for the string "笑話" across all texts in the Chinese Text Project, grouped by work and ordered by date of authorship, listing the first occurrence within each text. The chart at the top shows the proportion of all texts per year containing the string, from 400 BC through 1900 AD. In this case, the trend of the chart demonstrates the emergence of "笑話" as a widely attested written compound from the Song dynasty onwards.

This simple example points the way toward more sophisticated non-traditional ways of analyzing properties of text digitally. Some of these techniques, such as vocabulary-based statistical analyses and stylometry, are relatively new and experimental; others, such as text reuse identification and pattern search using regular expressions, are more established and well understood. An important observation however is that no digital library, no matter how advanced or comprehensive its functionality, can hope to provide tools covering all possible use cases or emergent techniques. For this reason, instead of integrating as many text analysis tools as possible directly into the system, an Application Programming Interface (API) is provided, which allows third parties to extract text and data from the system dynamically as needed from within their own software.[6] This mechanism also allows direct connections to be made between the user interfaces of independently created systems – as is the case with the MARKUS markup tool, developed by Hilde De Weerdt and Brent Ho,[7] which can be directly accessed from within the Chinese Text Project interface to annotate textual materials contained in the system.

While the API can be used to facilitate specialized text mining as part of research projects involving the construction of custom software, it can also enable open and decentralized deployment of analytic tools accessible to much wider audiences with much lower barriers to entry. One example of this is Text Tools,[8] which provides a point and click interface via a web browser to perform a variety of text mining tasks, including pattern search with regular expressions, comparisons of similarity of vocabulary and wording, together with a range of associated interactive visualizations. Running in the user's web browser, this program allows a user to perform a variety of tasks using materials from the Chinese Text Project and/or other sources. A concrete example of its use for text reuse identification and visualization is shown in Figure 3. In this case, the comparison is between chapters of the 史記, visualized in tabular form as a heat map; cells in the table are shaded darker to indicate higher levels of

[6] Donald Sturgeon, "Chinese Text Project: a dynamic digital library of premodern Chinese," *Digital Scholarship in the Humanities* (2019, Advance articles).

[7] Hilde De Weerdt, Ming-kin Chu, and Hou-ieong Ho, "Chinese Empires in Comparative Perspective: A Digital Approach," *Verge: Studies in Global Asias* 2.2 (2016), 58–69.

[8] Donald Sturgeon, "Digital Approaches to Text Reuse in the Early Chinese Corpus," *Journal of Chinese Literature and Culture* 5.2 (2018), 186–213.

reused text between each pair of chapters. This simple idea, repeated across all of the chapters of the Annals and Hereditary Houses sections, shows at a glance where material appearing in the Annals is repeated in a biography. Once the logic through which the visualization is created is understood, complex patterns can be grasped intuitively from the picture: the shaded region along the diagonal from the top-left of the table indicates that successive chapters of the Annals tend to repeat in identical or very similar words material from the previous chapter; biographies repeating in similar form the contents of specific Annals are highlighted by the isolated dots on the right hand side of the table – such as the biography of King Daohui of Qi, which contains many extended near-identical passages to the Annal of Empress Lü. The visualization takes only moments to produce, and when viewed within the tool allows immediate introspection of all of the precise textual details which contribute to the visual summary.

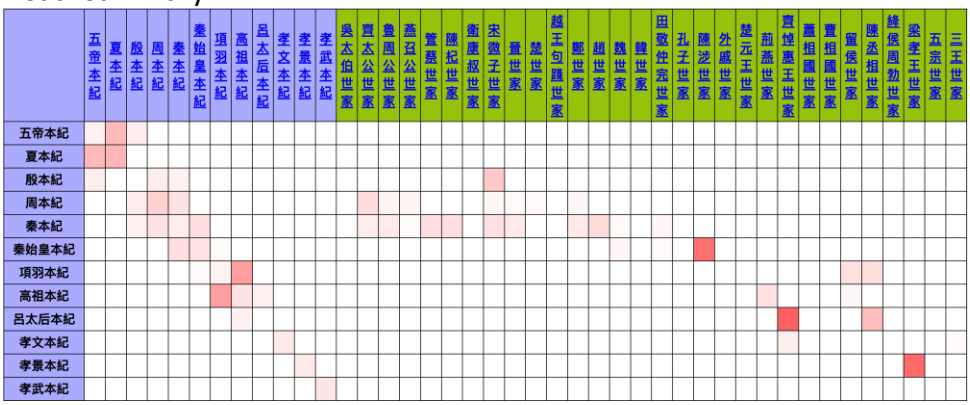


Figure 3: Text reuse in chapters of the Shiji. Blue rows and columns represent chapters of the Annals (本紀), green columns chapters of the Hereditary Houses (世家).

Annotation

The widespread availability of digitally transcribed text has transformed entirely the range of tasks that can be productively accomplished with historical materials by computer software. Particularly as compared with digital images of primary source material, digital transcriptions are transformative in the processing they facilitate computationally. Yet at the same time, transcription alone provides direct access to the semantic content of these materials only on a very superficial level: it is easy to determine which sequences of characters are identical and which are not, and to use this information to investigate related formal properties such as text reuse, but it remains difficult to draw inferences about the meaning of the contents with much certainty. In the case of Chinese, the lack of explicit delimiters such as spaces between words makes this even more apparent: even the seemingly trivial task of listing all the distinct words contained in a text in literary Chinese remains a challenging task for a computer when given only a transcription of its contents. While Natural Language Processing (NLP) techniques – such as the automatic identification of word boundaries, parts of speech of individual words, and generation of complete parse trees for individual sentences – have improved immensely through many decades of research and development, many of these techniques have not yet been adapted to work satisfactorily with premodern Chinese sources generally. Moreover, even when successful these techniques come with a non-trivial error rate, much like OCR, meaning that completely accurate results cannot be expected.

Annotation provides an intuitive way of enriching the textual representation so that it can provide computer software with additional information about not just what a text *says*, but also capture

precisely some aspects of what it *means*. To give a simple example, faced with the short sentence “孟子見梁惠王。” (Mengzi went to see King Hui of Liang), absent of any prior knowledge about the sentence, to a computer this simply represents a sequence of characters, “孟”, “子”, “見”, etc. Natural language processing techniques, such as tokenization and named entity detection, might be able to transform this into a sequence of words – “孟子”, “見”, and “梁惠王” – while attaching to this sentence the knowledge that “孟子” and “梁惠王” are proper names referring to people; where such techniques remain inadequate, human readers can perform this task instead. One way of representing this information is using a markup language, which intersperses machine-readable codes – “markup” – with the text itself. The most widely used language for this is XML (eXtensible Markup Language), in which markup consists of paired codes or “tags” surrounding the regions of text to which they apply. In this simple example, we might encode the knowledge that – in this particular sentence – “孟子” and “梁惠王” are names of people using the following code:

```
<person>孟子</person>見<person>梁惠王</person>。
```

In this example, the markup makes explicit the claim that “孟子” is the name of a person, because it falls between the opening “<person>” and closing “</person>” tags; the same is true of “梁惠王”, and nothing in particular is claimed about “見”. While expressing information trivial to a human reader familiar with the language, this encoded form of the text immediately provides computer software with far better information about what the text means than a plain transcription of the same sentence. Given an entire text – or large corpus of texts – marked up with codes like these, a search or statistical analysis tool can trivially make use of information about which names are referred to in the text, without having to perform any complex statistical estimation or incur any additional error rate in performing this task.

Crucially, this approach also allows for attaching additional data to specific parts of the text to further supplement the information available for subsequent computer processing. Keeping with this example, we might want to encode information about *people* rather than simply *names*. When we move beyond trivial examples of one or two sentences and start looking at larger scales, it becomes useful to be able group together proper names which refer to the same individual – “孟子” and “孟軻”, or “梁惠王” and “文惠君” – rather than simply those names that are identical. At the same time, distinguishing between different individuals who could be referred to by the *same* name necessitates recording some information to indicate which names refer to which people – e.g. to record the information that the 孟子 in “孟子見梁惠王” is the same person as the 孟軻 of “孟軻困於齊梁”, while also being different from the 孟子 in “惠公元妃孟子”. Identifiers (often abbreviated to “ID”) provide a simple way of achieving this: two names are given the same identifier if and only if they refer to the same person. In XML, this information can be added to a span of text by including it in the left-side tag:

```
<person id="1">孟子</person>見<person id="2">梁惠王</person>。
<person id="1">孟軻</person>困於齊梁。
<person id="3">惠公</person>元妃<person id="4">孟子</person>。
```

The same idea extends naturally to accommodate other semantic information present in text but otherwise inaccessible to computational processing. Examples include place names, era names, dates, bureaucratic office titles, and potentially many other types of data, up to and including semantic and grammatical roles of all individual words. Once this information is encoded, it offers the prospect not only for more advanced digital library functionality, but also large-scale computer-assisted statistical analyses making use of this additional machine-readable information. Identifiers used to distinguish the referents can then be used to link information across different types of digital system. In the case of person names, databases containing extensive information about historical individuals disambiguated using identifiers in precisely this way already exist in a number of domains – library systems such as the Library of Congress being one example, and in the historical Chinese domain scholarly research databases such as the China Biographical Database (CBDB)[9] and the Buddhist Studies Authority Database Project.[10] The significance of connecting these identifiers to identifiers in textual markup is that once linked in this way information from any or all of the linked resources can be pooled together and questions asked of the combined data. For example, given a text containing references to many thousands of historical individuals and annotated with identifiers linking them to historical persons in CBDB, any properties contained in CBDB can then be used as a dimension of analysis – which regions most of these people were associated with, which people held the highest offices, or which people shared kinship or social relationships of various kinds. Because the identifiers make explicit which person is referred to in each case, putting aside errors in the datasets and sources themselves, such queries can then be answered precisely without resorting to fallible estimation through NLP. This conceptual linking of resources using shared identifiers – often referred to as Linked Open Data – results in a growing network of machine-readable semantic relationships between independently maintained and operated projects.

Crucially, for the foreseeable future much of this semantic information will rely at least in part on human intervention or review for reliable creation. Context and domain knowledge are necessary in determining the true referents of terms, and for many historical cases – names of countless individuals less well known or attested in the historical record than Mengzi, say – expert judgement will be necessary. These observations have led to the creation of efficient semi-automated annotation systems such as Recogito[11] and Markus,[12] which assist a human annotator in the task of taking an unannotated text, efficiently identifying regions of text which should be marked up, and creating machine-readable annotations connecting them to their referents or incorporating other relevant information. These tools are often intended to be used in workflows where a researcher spends time marking up a corpus of texts as a first step to further analysis based on the marked-up content.

In the Chinese Text Project, annotation serves closely related purposes, but operates in a subtly different problem space. Like Recogito and Markus, a key goal is to facilitate semi-automated mark-up of content and enable linking of content to external sources of data. Unlike these tools, the Chinese Text Project also has the goal of layering at least some of this markup on top of publicly accessible

---

[9] https://projects.iq.harvard.edu/cbdb/home

[10] https://authority.dila.edu.tw/

[11] Rainer Simon, Elton Barker, Leif Isaksen, and Pau de Soto Cañamares, “Linking Early Geospatial Documents, One Place at a Time: Annotation of Geographic Documents with Recogito,” *e-Perimetron* 10.2 (2015), 49–59.

[12] De Weerdt et al, “Chinese Empires in Comparative Perspective”.

transcriptions which also remain tightly linked to the primary sources, and additionally operates on the assumption that many individuals will contribute independently over time to the overall annotation project. This necessitates a degree of coordination as to what information will be marked up, and how – something less of a concern when a single researcher is preparing data for individual use. The promise of this shared-text approach is that as annotations contributed by many individual users accumulate over time, a body of pre-prepared marked-up texts will become available to anyone wishing to perform analyses using this shared data. Rather than each researcher having to identify and disambiguate each annotation before being able to use annotations in any analysis, the data can be stored with the public copy of the text and downloaded or delivered to external tools via API as needed.

In order to make this approach work, for referring terms a database of entities referred to is also required. This models the intuition that there are many facts about individuals (and other entities such as places or eras) that it will be important for the annotation system itself to be aware of, but which we would not want to have to add in our markup every time a reference to that entity is made – for instance, the fact that "孟子" and "孟軻" can both be used to refer to the same person, and that this person was alive within some particular window of time. In the Chinese Text Project, this database is implemented as a crowdsourced entity graph database, recording statements about entities – and, where possible, providing textual sources for individual claims. With this basic infrastructure in place, the annotation workflow consists of taking a transcribed text (potentially containing some preexisting annotations), identifying possible matches with entities according to the data in the entity database, and then deciding – using a combination of automated and manual work – which of these possible candidates do in fact refer to which entities (and/or new entities attested in this text that need to be created in the entity database), and finally committing these changes to the public copy of the text. The changes are versioned through the same crowdsourcing system used to manage changes to the transcription itself, and the process can be repeated by the same user or any other user in future to make corrections or to add instances that were missed. The machine-readable data produced through this workflow can be used to enhance the user interface and search functionality, but also to provide direct access to the annotations via API.

Aside from referring terms, another type of textual reference that benefits greatly from markup is explicit reference to dates. In the simplest case, dates may consist of a precise and complete description specifying a day indexed according to a particular era, year and month of rule, and day in the 60-day cycle – as in a statement like "元祐元年春正月庚寅朔，改元。". While ultimately the corresponding date in a modern calendar system will be useful for reference and analysis, in the Chinese Text Project implementation only the historically attested information necessary to *interpret* this date – that it represents day 庚寅 of the first month of the first year of the Yuanyou era – is encoded using markup, with reference to an entity representing the Yuanyou era. The same is done for "partial" dates, in which the statement itself does not contain all of the information necessary to interpret the date, yet the context in which it appears nevertheless makes it entirely unambiguous – as in "癸巳，王安石薨。", the context of which in its location in the Songshi (宋史) makes quite clear that this means the fourth month of the first year of the Yuanyou era. In this latter case, the annotation process adds this contextual information to the markup, meaning that this instance of "癸巳" can be treated quite differently from the superficially identical occurrence in the phrase "癸巳，謝奕昌卒。".

What is intentionally *not* encoded is the correspondence between this date and any modern calendar – this is treated as an additional interpretative step which is taken using a separately implemented model of how these mappings should be made. This contrasts with most other implementations in which the literal contextualized content of the date is not recorded in the markup, but instead a date in standardized Julian or Gregorian calendar format is stored. Separating these two tasks greatly simplifies the markup task: editors need only decide which year, month, and day of which era is intended by the reference; the less straightforward task of deciding which Julian or Gregorian day that corresponds to is left to computer software, which – armed with an interpretative model, currently provided by the Buddhist Studies Time Authority Database[13] – can perform these calculations unambiguously on the marked-up text. Thus, the marked-up dates become precise, machine-readable references, which can be used within the library to display to the user dates in a Western calendar, or to locate references to dates within particular ranges regardless of how they are expressed, as well as being exportable for larger statistical analyses.

Markup, entities, and date handling together also facilitate crowdsourced editing of claims about entities. For instance, the statement "癸巳，王安石薨。", now with machine-readable time context, can be used as evidence for a claim about this individual: that he died on the day 癸巳 of the fourth month of the first year of the Yuanyou era. This claim is recorded in the entity record for Wang Anshi, together with the precise textual source of the claim (Table 1). Over time, further information can be added, producing a fully sourced and annotated record of the primary source textual basis for claims about individuals. As with markup of texts, in general these claims often require human input to accurately create, but in some cases they can be reliably inferred from context and other annotations – for example, directly analogous cases such as "庚戌，孟昶薨。" and "壬午，張知白薨。". Provided that dates and proper names have been marked up first, these claims can be reliably and mechanically inferred from the text, and added – with citations – to the appropriate entity records. Alongside these claims based on primary source data, identifiers from external sources and databases uniquely identifying the same entity are included – for example, CBDB identifiers and Dharma Drum identifiers for historical people.

Relation	**Value**	**Textual reference**
type	person	
name	王安石	
name-style	介甫	《宋史·列傳第八十六》：王安石，字介甫，撫州臨川人。
jiguan	place:臨川	《宋史·列傳第八十六》：王安石，字介甫，撫州臨川人。
father	person:王益	《宋史·列傳第八十六》：父益，都官員外郎。
died	元祐元年四月癸巳	《宋史·本紀第十七》：癸巳，王安石薨。
authority-cbdb	1762	
authority-ddbc	A007519	

Table 1: Fragment of the entity record for Wang Anshi. All relations, statements, dates, associations with other entities, and textual references in the entity record are machine-readable and can be processed automatically without scope for ambiguity. The last two lines provide explicit links to locate the same entity in the China Biographical Database and Buddhist Studies Person Authority Database.

[13] https://authority.dila.edu.tw/time/

In a similar way to crowdsourced transcription, this approach facilitates access to another "long tail": while many claims, such as the date of death of well-known historical figures, will be contained in mainstream sources such as the Songshi, others may be sparsely distributed across a range of sources. Note that even in the simplest case of a famous individual such as Wang Anshi with his own biography in the Songshi, the claim of a precise date of death is not contained in the biography itself – which states only 元祐元年 – but instead in the annals of Emperor Zhezong. Tying these claims precisely to their textual sources makes their attestations easily verifiable in a crowdsourced editing context.[14] Furthermore, as with the markup task, while human input is required in the general case, some knowledge can nevertheless be reliably extracted using automated methods – many of which have been successfully demonstrated during the construction of CBDB.[15] Finally, the marked-up sources – now carrying machine-readable information precisely disambiguating references to dates and people – can be further mined to extract additional knowledge, for inclusion into the entity database as well as reuse in other projects.

The machine-readable knowledge claims created through this process can also be reused by the system itself in the future markup of texts. For example, a reference to a person by the name of "王信" without context is ambiguous – even just within the Songshi there are two biographies of individuals with that name – but if prefixed by the title "給事中", a likely candidate can be selected given the knowledge that one person by such a name held this particular title. Similar types of inference can be made based on data relating to dates, as well as much more complex types of analyses using combinations of data points.

Conclusions and future work

The enormous growth in computational power and storage over the past few decades has transformed what can be practically achieved with computational processing of literary sources. Together with the spread of the internet, this has reduced by many orders of magnitude the time and effort required to access primary source materials and locate information within them. Adoption of web browsers adhering to internet standards has led to the creation of digital platforms allowing immediate access to a wealth of sources and tools for working with them that would previously have been unimaginable. However, despite many important and exciting advances in natural language processing, computational processing of textual data is frequently limited in capabilities and accuracy by the complexity of natural language. Human input, through annotation and the creation of machine-readable datasets, offers the immediate prospect of digital platforms that can reliably make more complex inferences from historical sources.

As these machine-readable, annotated sources accumulate, they will naturally form the basis for further development of automated techniques that can better approximate human annotation of text. Supervised machine learning – in which software learns patterns automatically from a set of examples, which it can then apply to unseen but analogous cases – has been extremely successful in

[14] "Verifiable" here means having the ability to confirm that such a claim was made in some particular primary source text – not, of course, that the claim itself is historical fact.

[15] Chao-Lin Liu, Chih-Kai Huang, Hongsu Wang, and Peter K. Bol, "Toward Algorithmic Discovery of Biographical Information in Local Gazetteers of Ancient China," *29th Pacific Asia Conference on Language, Information and Computation* (2015) 87–95.

similar tasks on other domains. The same body of annotated texts that is immediately useful to literary and historical scholars will also form the basis for advancements in its application to premodern Chinese materials.

At the same time, the importance of improved metadata and expanded linking and data exchange between projects is becoming clear. Machine-readable identifiers together with APIs have been shown to be invaluable in achieving scalable, sustainable, and meaningful connections between independently developed and maintained digital systems. Better standardization and interoperability between these systems will be needed in order to develop distributed models of working, in which the countless as-yet uninvestigated possibilities offered by the digital medium can be fruitfully explored.